RMarkdown STAT 2218 Project 1

Jack Ambery and Kevin Wilson

Nov 3, 2023

Work in a team of 1, 2 or 3 students. You will hand in 1 typed report per group.

You are hired by a company sells two main types of products remotely by connecting to and talking to customers by phone. They provide a data set that has the following information. Each row includes information from one employee’s shift randomly selected across the different teams, regions and products. Read through the variables carefully.

Data: salesdata.csv

* Team: This represents the team that the employee works on (Team A to J). The company is organized into 10 teams.
* Product: Which product was being sold during shift (Product 1 or 2)
* Region: Which state the employee works in
* Shift: The number of hours the employee worked on that shift
* TotalTalk: The total number of hours during the shift the employee was talking to customers on the phone
* TotalHold: The total number of hours during the shift the employee had customers on hold
* TotalNumSales: The total number of items sold during the shift
* TotalRevenue: The total amount of revenue (in $) employee made during the shift
* NumCalls: The total number of calls the employee took during the shift
* TotalNumFeedback: The total number of customers that submitted feedback on the employee during the shift
* PercentSatisfied: The percent of customer surveys that had positive reviews (Excellent or Good ratings)
* NumPos: The number of customer surveys that were positive reviews (Excellent or Good ratings)
* NumNeg: The number of customer surveys that were negative reviews (Fair or Poor ratings)

For this project, they want you to generate a linear model that predicts total revenue from other metrics that they collect. Your team will write a report based on your analyses, assuming a significance level of 5%, that includes your R code, relevant output and analysis for each of following concepts.

# Input data file below.

Import data into RStudio using the base option, and then cut and paste YOUR file directory in the read.csv command below. It may be different from what I have listed below since your files are located in a different directory/folder. Then you can attach the file.

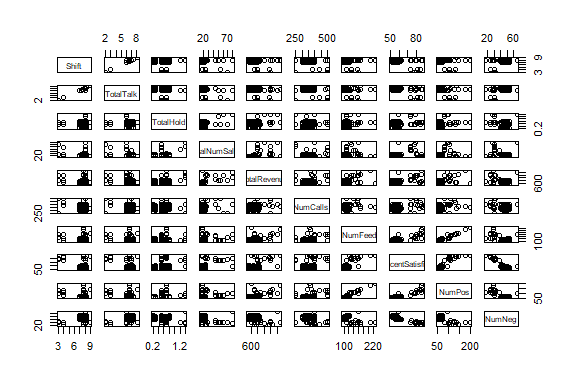
salesdata <- read.csv("C:/Users/jacka/OneDrive - fairfield.edu/Fall 2023/Statistics II/salesdata.csv")  
attach(salesdata)

# Part 1: Exploratory Analysis (10 pts)

* What numerical explanatory variables appear to be related to total revenue? Are they linearly related? Explain and provide graphical and numerical evidence.
* What numerical explanatory variables appear to be related to each other? Explain and provide graphical and numerical evidence.

## Part 1 R Code

# Plot of all the numerical values plotted against each other  
plot(salesdata[, 4:13])



#Correlations of each variable pair combination  
cor(salesdata[, 4:13])

## Shift TotalTalk TotalHold TotalNumSales TotalRevenue  
## Shift 1.0000000 0.98068241 0.02661050 -0.1927238 -0.1701240  
## TotalTalk 0.9806824 1.00000000 -0.08776289 -0.3346494 -0.2437727  
## TotalHold 0.0266105 -0.08776289 1.00000000 0.4508702 0.3511603  
## TotalNumSales -0.1927238 -0.33464936 0.45087022 1.0000000 0.6055547  
## TotalRevenue -0.1701240 -0.24377267 0.35116034 0.6055547 1.0000000  
## NumCalls -0.2061829 -0.19595403 0.04478027 0.1215173 0.2793540  
## TotalNumFeedback -0.2174747 -0.23694242 0.19533074 0.2325233 0.5133673  
## PercentSatisfied -0.3523145 -0.39456923 0.24987222 0.4112045 0.5323060  
## NumPos -0.2787064 -0.30279980 0.21736512 0.2852604 0.5369614  
## NumNeg 0.3175196 0.34325302 -0.18154936 -0.2994607 -0.3692502  
## NumCalls TotalNumFeedback PercentSatisfied NumPos  
## Shift -0.20618292 -0.2174747 -0.3523145 -0.2787064  
## TotalTalk -0.19595403 -0.2369424 -0.3945692 -0.3027998  
## TotalHold 0.04478027 0.1953307 0.2498722 0.2173651  
## TotalNumSales 0.12151731 0.2325233 0.4112045 0.2852604  
## TotalRevenue 0.27935397 0.5133673 0.5323060 0.5369614  
## NumCalls 1.00000000 0.2584633 0.3279796 0.2946728  
## TotalNumFeedback 0.25846332 1.0000000 0.7562516 0.9585545  
## PercentSatisfied 0.32797964 0.7562516 1.0000000 0.9036830  
## NumPos 0.29467283 0.9585545 0.9036830 1.0000000  
## NumNeg -0.26240481 -0.4444557 -0.8982169 -0.6812569  
## NumNeg  
## Shift 0.3175196  
## TotalTalk 0.3432530  
## TotalHold -0.1815494  
## TotalNumSales -0.2994607  
## TotalRevenue -0.3692502  
## NumCalls -0.2624048  
## TotalNumFeedback -0.4444557  
## PercentSatisfied -0.8982169  
## NumPos -0.6812569  
## NumNeg 1.0000000

## Part 1 Analysis

Insert your analysis here.

Based on the table of scatterplots we created, it initially does not appear that many of the variables follow a linear trend that may predict Total Revenue. Based on the scatter plots, it seems that the only variables that have the potential to predict TotalRevenue are TotalNumSales, NumPos, and TotalNumFeedback. Further analysis will be required to confirm this.

Looking at the correlation coefficients, it appears that TotalNumSales, TotalNumFeedback, PercentSatisfied, and NumPos have the strongest relationships with TotalRevenue. These have correlations of 0.61, 0.51, 0.53, and 0.54, respectively. This confirms the slight trends we noticed in the scatterplots.

Next, we looked at linear relationships between explanatory variables. We noticed quite a few on the scatterplot table and confirmed those observations with the correlation coefficients.

Shift and TotalTalk have a correlation coefficient of about 0.98, there is definitely a linear relationship between Shift and TotalTalk. Intuitively, this makes a lot of sense.

TotalNumFeedback and NumPos appear to be linearly related as they have a correlation coefficient of 0.96.

PercentSatisfied and NumPos appear to be linearly related as they have a correlation coefficient of 0.90.

PercentSatisfied and NumNeg appear to have a linear relationship as they have a correlation coefficient of -0.90.

PercentSatisfied and TotalNumFeedback have a correlation coefficient of 0.756, suggesting a moderately strong correlation.

NumPos and NumNeg appear to have a linear relationship as they have a correlation coefficient of -0.68.

These relationships will help us determine possible multicollinearity later, when trying to deteremine the best linear model to use.

# Part 2: Models (30 pts)

* Find the model that your group best predicts total revenue from all or a subset of the numerical explanatory variables.
* You should summarize all models that your group looked at and explain why you judged your final model to be best.
* Assess your final model, including checking for multicollinearity and determine if each condition of the multiple linear regression (MLR) assumptions is met. Explain and include evidence to support your claims.
* If your final model exhibits multicollinearity, find and include a different model that does not exhibit multicollinearity. And, explain which model (the model with multicollinearity or without multicollinearity) you think is better and why.
* Give the equation of your group’s “best” MLR model to predict total revenue.

## Part 2 R Code

# This is the linear model that uses every numerical variable to predict TotalRevenue  
out=lm(TotalRevenue ~ Shift + TotalTalk + TotalHold + TotalNumSales + NumCalls + TotalNumFeedback + PercentSatisfied + NumPos + NumNeg)  
summary(out)

##   
## Call:  
## lm(formula = TotalRevenue ~ Shift + TotalTalk + TotalHold + TotalNumSales +   
## NumCalls + TotalNumFeedback + PercentSatisfied + NumPos +   
## NumNeg)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -506.11 -86.27 -6.17 75.66 469.88   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 315.7964 668.4048 0.472 0.6375   
## Shift -99.9200 111.1429 -0.899 0.3706   
## TotalTalk 109.6314 114.2919 0.959 0.3395   
## TotalHold 62.5934 60.1266 1.041 0.3001   
## TotalNumSales 9.2838 1.7839 5.204 9.01e-07 \*\*\*  
## NumCalls 0.5380 0.3123 1.723 0.0878 .   
## TotalNumFeedback -0.1888 5.7211 -0.033 0.9737   
## PercentSatisfied -4.4701 9.2264 -0.484 0.6290   
## NumPos 4.3024 7.2841 0.591 0.5560   
## NumNeg NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 130.8 on 111 degrees of freedom  
## Multiple R-squared: 0.5377, Adjusted R-squared: 0.5044   
## F-statistic: 16.14 on 8 and 111 DF, p-value: 1.3e-15

# removed TotalNumFeedback  
out=lm(TotalRevenue ~ Shift + TotalTalk + TotalHold + TotalNumSales + NumCalls + PercentSatisfied + NumNeg + NumPos)  
summary(out)

##   
## Call:  
## lm(formula = TotalRevenue ~ Shift + TotalTalk + TotalHold + TotalNumSales +   
## NumCalls + PercentSatisfied + NumNeg + NumPos)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -506.11 -86.27 -6.17 75.66 469.88   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 315.7964 668.4048 0.472 0.6375   
## Shift -99.9200 111.1429 -0.899 0.3706   
## TotalTalk 109.6314 114.2919 0.959 0.3395   
## TotalHold 62.5934 60.1266 1.041 0.3001   
## TotalNumSales 9.2838 1.7839 5.204 9.01e-07 \*\*\*  
## NumCalls 0.5380 0.3123 1.723 0.0878 .   
## PercentSatisfied -4.4701 9.2264 -0.484 0.6290   
## NumNeg -0.1888 5.7211 -0.033 0.9737   
## NumPos 4.1136 1.8952 2.170 0.0321 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 130.8 on 111 degrees of freedom  
## Multiple R-squared: 0.5377, Adjusted R-squared: 0.5044   
## F-statistic: 16.14 on 8 and 111 DF, p-value: 1.3e-15

# removed NumNeg  
out=lm(TotalRevenue ~ Shift + TotalTalk + TotalHold + TotalNumSales + NumCalls + PercentSatisfied + NumPos)  
summary(out)

##   
## Call:  
## lm(formula = TotalRevenue ~ Shift + TotalTalk + TotalHold + TotalNumSales +   
## NumCalls + PercentSatisfied + NumPos)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -504.75 -86.48 -6.28 75.56 470.53   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 294.9852 220.8360 1.336 0.184334   
## Shift -100.4205 109.6117 -0.916 0.361558   
## TotalTalk 110.1836 112.5554 0.979 0.329725   
## TotalHold 62.7460 59.6807 1.051 0.295356   
## TotalNumSales 9.2681 1.7117 5.415 3.55e-07 \*\*\*  
## NumCalls 0.5365 0.3075 1.745 0.083760 .   
## PercentSatisfied -4.1890 3.5334 -1.186 0.238314   
## NumPos 4.0653 1.1999 3.388 0.000973 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 130.2 on 112 degrees of freedom  
## Multiple R-squared: 0.5377, Adjusted R-squared: 0.5088   
## F-statistic: 18.61 on 7 and 112 DF, p-value: 2.879e-16

# removed Shift  
out=lm(TotalRevenue ~ TotalTalk + TotalHold + TotalNumSales + NumCalls + PercentSatisfied + NumPos)  
summary(out)

##   
## Call:  
## lm(formula = TotalRevenue ~ TotalTalk + TotalHold + TotalNumSales +   
## NumCalls + PercentSatisfied + NumPos)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -510.72 -89.25 -2.42 76.73 465.67   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 223.1115 206.2842 1.082 0.281745   
## TotalTalk 7.9295 14.5224 0.546 0.586129   
## TotalHold 39.4894 53.9744 0.732 0.465907   
## TotalNumSales 8.2547 1.3053 6.324 5.26e-09 \*\*\*  
## NumCalls 0.5828 0.3031 1.923 0.056976 .   
## PercentSatisfied -4.2899 3.5292 -1.216 0.226696   
## NumPos 4.1475 1.1957 3.469 0.000741 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 130.1 on 113 degrees of freedom  
## Multiple R-squared: 0.5343, Adjusted R-squared: 0.5095   
## F-statistic: 21.6 on 6 and 113 DF, p-value: < 2.2e-16

# removed TotalTalk  
out=lm(TotalRevenue ~ TotalHold + TotalNumSales + NumCalls + PercentSatisfied + NumPos)  
summary(out)

##   
## Call:  
## lm(formula = TotalRevenue ~ TotalHold + TotalNumSales + NumCalls +   
## PercentSatisfied + NumPos)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -507.60 -90.69 -3.34 77.34 469.76   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 300.2968 149.7718 2.005 0.047329 \*   
## TotalHold 42.2823 53.5658 0.789 0.431545   
## TotalNumSales 8.1082 1.2735 6.367 4.19e-09 \*\*\*  
## NumCalls 0.5699 0.3012 1.892 0.061009 .   
## PercentSatisfied -4.6828 3.4444 -1.360 0.176659   
## NumPos 4.2052 1.1874 3.542 0.000577 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 129.7 on 114 degrees of freedom  
## Multiple R-squared: 0.533, Adjusted R-squared: 0.5126   
## F-statistic: 26.03 on 5 and 114 DF, p-value: < 2.2e-16

# removed TotalHold  
out=lm(TotalRevenue ~ TotalNumSales + NumCalls + PercentSatisfied + NumPos)  
summary(out)

##   
## Call:  
## lm(formula = TotalRevenue ~ TotalNumSales + NumCalls + PercentSatisfied +   
## NumPos)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -506.68 -87.99 2.62 75.92 472.17   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 317.3563 147.9612 2.145 0.034068 \*   
## TotalNumSales 8.5099 1.1655 7.301 3.99e-11 \*\*\*  
## NumCalls 0.5611 0.3005 1.867 0.064419 .   
## PercentSatisfied -4.7559 3.4375 -1.384 0.169186   
## NumPos 4.2749 1.1821 3.616 0.000445 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 129.5 on 115 degrees of freedom  
## Multiple R-squared: 0.5305, Adjusted R-squared: 0.5142   
## F-statistic: 32.48 on 4 and 115 DF, p-value: < 2.2e-16

# removed PercentSatisfied  
out=lm(TotalRevenue ~ TotalNumSales + NumCalls + NumPos)  
summary(out)

##   
## Call:  
## lm(formula = TotalRevenue ~ TotalNumSales + NumCalls + NumPos)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -450.72 -93.08 4.90 78.81 502.93   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 170.1883 103.2512 1.648 0.1020   
## TotalNumSales 7.9097 1.0860 7.283 4.24e-11 \*\*\*  
## NumCalls 0.5003 0.2984 1.676 0.0963 .   
## NumPos 2.8168 0.5375 5.240 7.27e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 130 on 116 degrees of freedom  
## Multiple R-squared: 0.5227, Adjusted R-squared: 0.5103   
## F-statistic: 42.34 on 3 and 116 DF, p-value: < 2.2e-16

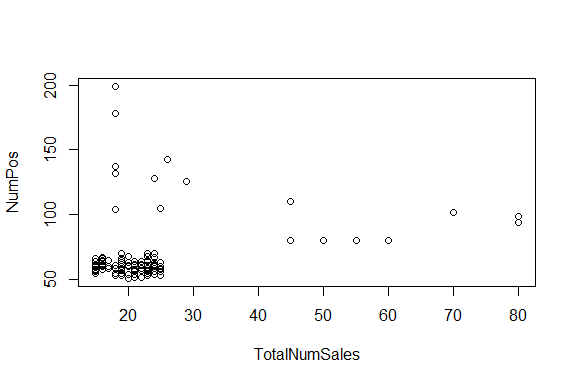
#Removed NumCalls  
out=lm(TotalRevenue ~ TotalNumSales + NumPos)  
summary(out)

##   
## Call:  
## lm(formula = TotalRevenue ~ TotalNumSales + NumPos)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -483.89 -101.21 -4.03 81.29 480.44   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 330.620 39.066 8.463 8.67e-14 \*\*\*  
## TotalNumSales 7.984 1.093 7.302 3.73e-11 \*\*\*  
## NumPos 3.063 0.521 5.879 3.99e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 131 on 117 degrees of freedom  
## Multiple R-squared: 0.5111, Adjusted R-squared: 0.5027   
## F-statistic: 61.16 on 2 and 117 DF, p-value: < 2.2e-16

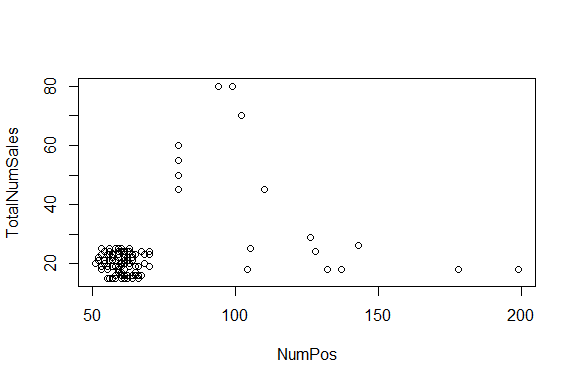
# this is our final model, all variables are significant  
  
# Checking for multicollinearity  
cor(TotalNumSales, NumPos)

## [1] 0.2852604

plot(TotalNumSales, NumPos)



plot(NumPos, TotalNumSales)



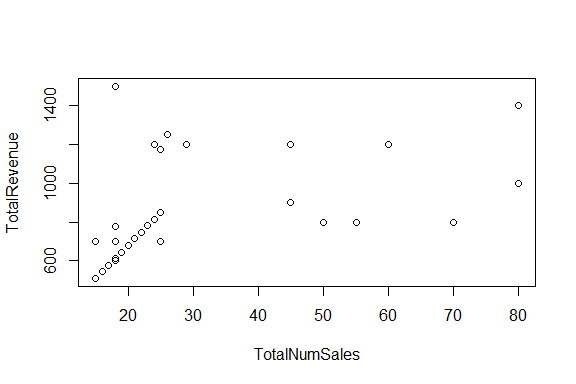
library(car)

## Loading required package: carData

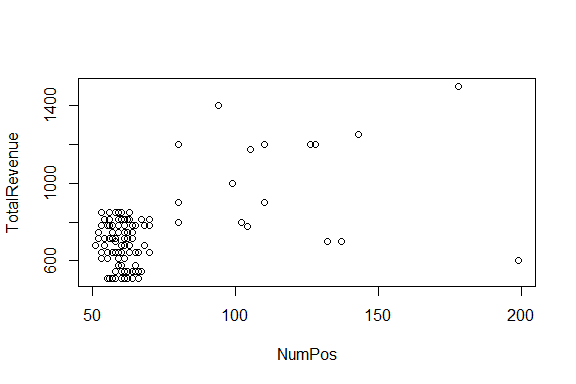
vif(out)

## TotalNumSales NumPos   
## 1.088582 1.088582

# Checking Assuptions  
# Assumption 1 - Linear model is appropriate?  
plot(TotalRevenue~TotalNumSales)



plot(TotalRevenue~NumPos)



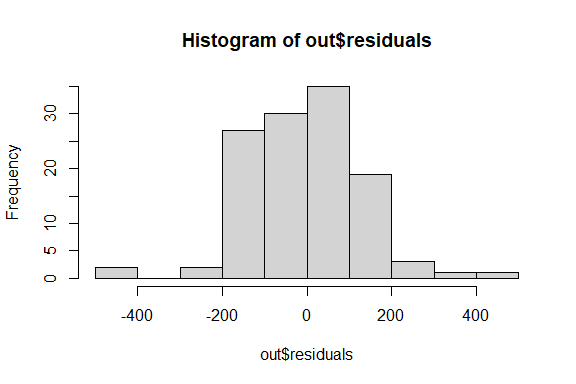
cor(TotalRevenue,TotalNumSales)

## [1] 0.6055547

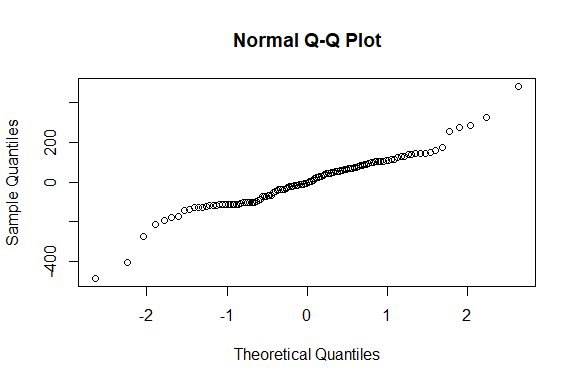
cor(TotalRevenue,NumPos)

## [1] 0.5369614

# Assumption 2 - Residuals are normally distributed?  
hist(out$residuals)



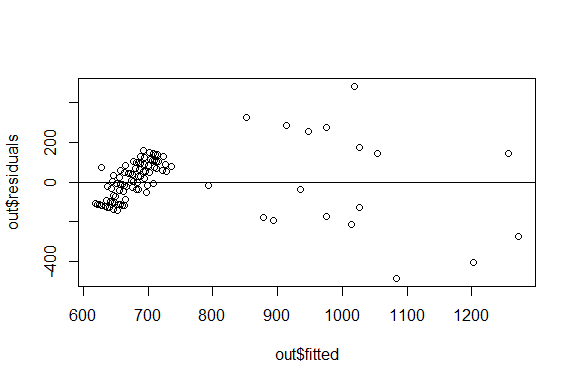
qqnorm(out$residuals)



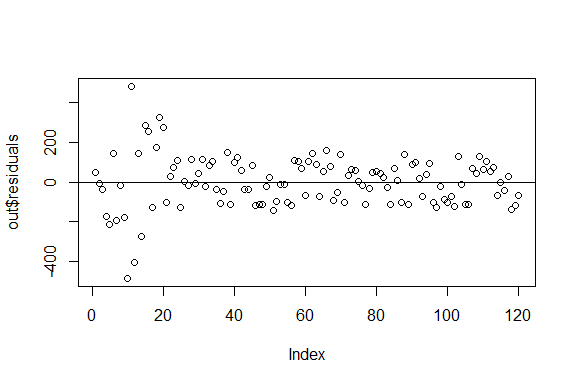
shapiro.test(out$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: out$residuals  
## W = 0.95571, p-value = 0.0005811

# Assumption 3 - Residuls has consant variance?  
plot(out$fitted, out$residuals)  
abline(0,0)



# Assumption 4 - Residuals are independent?  
plot(out$residuals)  
abline(0,0)



## Part 2 Analysis

The model including every numerical variable was not a significant in predicting TotalRevenue, since many variables had a p-value > 0.05 Also, the p-value for whole model was less than 0.05, so we do believe there is a linear relationship, we just need to eliminate variables that do not assist in this prediction.

Using the backwards elimintion method, we started removing variables from our linear model. First, we removed TotalNumFeedback from the next model because it had the highest p-value. This new model still had insigificant p-values, namely Shift, TotalTalk, TotalHold, NumCalls, PercentSatisfied and NumNeg. Through an iterative process, we kept removing the variable with the highest p-value one-by-one. The largest p-value in this model was NumNeg so we remove that variable for our next model.

Then the model still had insignificant p-values with the largest being Shift, so we removed Shift for our model.

This model also displayed insigificant p-values with the largest value being TotalTalk, so we removed it from our model.

Next, we deleted TotalHold. We were getting closer since the model now only had two variables with insigificant p-values which were NumCalls and PercentSatisfied. We removed PercentSatisfied for out next model as it had the largest p-value. Ad then we removed NumCalls after that.

Finally, our model only ended up including TotalNumSales and NumPos, which both had significant p-values, meaning this was our final model and we were finished with backwards elimination.

Next we had to check for possible mulitcolliniarity between our explanatory variables.

The correlation coefficient between TotalNumSales and NumPos is about 0.285 which indicated weak correlation between TotalNumSales and NumPos. In addition, the pairwise scatterplots between the two variables do not indicate any kind of relationship.

We also looked at the VIF for each variable. The results above show that the VIFs for TotalNumSales and NumPos are both less than 5, which supports no multicollinearity. We can confidently say, these three findings indicate that our final model does not exhibit any multicollinearity.

Next we checked all the assumptions of our linear model.

The first of these was to check if a linear model is appropriate. We looked at the pairwise scatterplots and the pairwise correlations with Total Revenue as Y and TotalNumSales and NumPos as the X variables. Both plots display a moderate linear relationship. To confirm this, both pairwise correlations are greater than 0.5. As a result, it appears assumption 1 is met, a linear model is appropriate.

The next condition checks if the residuals are normally distributed and centered about 0. The histogram appears relatively normal and centered at 0. Additionally, the shapiro test returns a p-value of 0.00058, indicating the residuals are normally distributed. Thus, condition 2 is met as the residuals are normally distributed and centered about 0.

The third condition checks if the residuals have constant variance. We look at the plot of residuals vs predicted values to see if this condition is met. Our plot does not display constant spread about y = 0 and it resembles a sideways V-shape. This means the residuals do not have constant variance and this condition is not met.

The last assumption checks if the residuals are independent, we look at the plot of the residuals. For the most part, there is no pattern and the plot shows random scatter but there are some weird values on the left that are much large than the others. We believe the residuals are still independent making this assumption met.

The equation for our final model is TotalRevenue “hat” = 330.620 + 7.984(TotalNumSales) + 3.063(NumPos)

## Part 3: Predictions (10 pts)

Use your group’s “best” model to find a 95% interval to predict the total revenue for each new observation with the measurements given in the table on the handout. Summarize the result in a sentence. Give the R code used.

Note that you may not need to use all of the inputs given in the table.

## Part 3 R code

# Subject One Prediction  
subject1 = data.frame(TotalNumSales=25, NumPos=61)  
predict(out, subject1, interval='confidence', level=.95)

## fit lwr upr  
## 1 717.0706 691.3744 742.7667

# Subject Two Prediction  
subject2 = data.frame(TotalNumSales=100, NumPos=75)  
predict(out, subject2, interval='confidence', level=.95)

## fit lwr upr  
## 1 1358.763 1191.783 1525.743

## Part 3 Analysis

With 95% confidence, we can say that the total revenue for Subject 1 will be between 691.37 and 742.77 dollars.

With 95% confidence, we can say that the total revenue for Subject 2 will be between 1191.78 and 1525.74 dollars.

End of Problem Set